



Kim, Y. (2015) "Designated communities": through the lens of the web. *International Journal of Digital Curation*, 10(1), pp. 184-195.

Copyright © 2015 The Author

This work is made available under the Creative Commons Attribution 2.0 License (CC BY 2.0)

Version: Published

<http://eprints.gla.ac.uk/103801/>

Deposited on: 10 March 2015

“Designated Communities”: Through the Lens of the Web

Yunhyong Kim
HATII, School of Humanities
University of Glasgow

Abstract

The notion of a “designated community” has always been a rather elusive concept across the digital curation landscape. This paper is an effort to revisit the concept to stir up new discussions in the area. More specifically, this study offers a perspective on designated communities through the lens of the web, powered by developments in the last ten years in social media. The research presents a multi-faceted analysis of communities based on HTML content from online web pages to propose heuristics for defining designated communities based on the technology they adopt, properties of knowledge organisation, and how they link to each other. This impacts the building of quantifiable models of designated communities, estimating curation risks associated to the community and further, refining approaches to preservation strategies that meet the needs of the community.

Received 10 October 2014 | *Accepted* 10 February 2015

Correspondence should be addressed to Yunhyong Kim, HATII, University of Glasgow, 11 University Garden, Glasgow, G12 8QH, Scotland. Email: yunhyong.kim@glasgow.ac.uk

An earlier version of this paper was presented at the 10th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: <http://www.ijdc.net/>

Copyright rests with the authors. This work is released under a Creative Commons Attribution (UK) Licence, version 2.0. For details please see <http://creativecommons.org/licenses/by/2.0/uk/>



Introduction

The notion of a “designated community” has always been an elusive concept within the digital preservation community. This term has come into wide scale use within the archival community since its use in the reference model for an Open Archival Information System (OAIS), published as an ISO standard in 2003 (ISO, 2003). The notion has propagated a high level understanding of the concept¹ without a full discussion of the technical complexity locked within, and identifying community needs has been largely left as part of the work flow for policy development². The aim of this paper is to raise a new perspective on the designated community and to stir the digital preservation community into action to re-investigate this concept.

Ten years ago and in the years thereafter, digital preservation research has largely focused on object oriented engineering solutions to curatorial problems, with a tendency to avoid delving too much into the complexity behind designated communities and their needs. During the same time period, we have experienced the advent of social media, which has caused the explosion of visible online communities (e.g. O'Callaghan et al., 2013; Massimi et al., 2014). This phenomenon has opened up a window into the workings of communities (e.g. Coulon, 2005; Lampe, Ellison and Steinfeld, 2008; Rowson, Broome and Jones, 2010) in a way that was not possible in the first stages of digital preservation research.

The research reported here proposes that content analysis of web pages can provide insight when profiling designated communities to understand complexities that might arise in digital preservation processes for weblogs (e.g. for example the range of different types of material that need considering in the target information community). This investigation is only a lightweight study to open up approaches to analyse “designated community” as a complex system.

The concept of *complexity* is mentioned in many disciplines to refer to a number of different phenomena. The complexity under scrutiny here is limited to that related to emerging community behaviours regarding the variety of platforms and file formats adopted to disseminate information, the range of terms and concepts adopted to organise information, and the links established to communicate information. The investigation employs, as a proxy, a quantitative analysis of HTML markup, user-generated categories and topic tags, and link network to generate target community profiles.

The study makes three contributions to current research landscape:

- It presents heuristics for a comparative study of different online communities, on multiple levels, especially, highlighting those related to blogging communities,
- The work opens up the first steps towards automated processes for deriving a quantified model of a “designated community” that can be used by the web archiving community, and
- The discussion highlights an approach for how the results in this paper on online communities can be generalised to be applied to the general digital curation context.

¹ For example, see definition from York University (2013).

² For example, see Alliance for Permanent Access (n.d).

Datasets

The analysis reported here was carried out on web home pages. The study was limited to home pages in order to make the study comparable across the datasets without noise propagated through page type variations. The analysis was performed using a fresh crawl of URLs from three datasets.³ The datasets are summarised in Table 1.

Table 1. Datasets used in the study. Right most column is included as a preliminary indication of the variation in the HTML source code.

Dataset	No. of URLs	No. of unique “<!DOCTYPE>” declarations
Spinn3r	223 145	80
ClueWeb09	214 952	1420
BF16Cat	31 690	122

The first of these datasets comprise the crawl of 223 145 URLs from the “Spinn3r” dataset⁴. The URLs of the Spinn3r dataset are collected as weblogs and social media websites and were valid links at the time the dataset was initially harvested, between August and October 2008.

The second dataset is a crawl of 214 952 URLs from the “ClueWeb09” dataset⁵. The ClueWeb09 dataset represents general web pages, and were valid links when the dataset was collected between January and February 2009.

The third dataset is “BF16Cat”, consisting of 31 690 URLs of home pages across blogs in eleven subject areas and non-blog in five sectors. The URLs for BF16Cat were either collected from blogrolls of individual bloggers in the corresponding area, or through blog searching portals such as Technorati⁶, mathblogging.org, scienceseeker.org, scienceblogging.org and Independent Fashion Bloggers (IFB)⁷. The sources for BF16Cat are shown in Table 2.

Table 2. Subcategories and/or domains associated to BF16Cat.

Type	Subcategory	Size	Source
Blogs	Computer Science (CS)	41	StackOverflow
	Information Technology (IT)	138	Technorati “IT” category search
	Entertainment (ET)	110	Technorati “Entertainment” category search
	Fashion (FA)	164	Independent Fashion Bloggers
	Game (GA)	7	University of Glasgow PhD student in Games
	Health Blogs (HB)	130	Technorati “Health” category Search

³ Crawl performed in July 2012.

⁴ ICWSM 2009 Spinn3r blog dataset: <http://www.icwsm.org/data/>

⁵ ClueWeb09 dataset: <http://www.lemurproject.org/clueweb09.php/>

⁶ Technorati: <http://technorati.com/>

⁷ Independent Fashion Bloggers: <http://heartifb.com>

Type	Subcategory	Size	Source
Non-blogs	Mathematics I (M1)	110	Field's Medalist Terry Tao's Blog
	Mathematics II (M2)	552	Mathblogging.org
	Music (MU)	70	Technorati "Music" category search
	Politics (PO)	107	Technorati "Politics" category search
	Science (SC)	1071	Scienceseeker.org and Scienceblogging.org
	Construction Company (CC)	27	National Building Specification website ⁸
	Fashion Company (FC)	61	Smashing Magaazine Showcase of Beautiful Fashion Websites ⁹ and comments
	Funding Council (FU)	51	Search on Google
	Government (GO)	572	Politics Resources website ¹⁰
	University (UN)	100	Sedghi and Evans (2012)

Examining HTML tags in the dataset, it was found that the total number of tags were 2 676 612, 139 176 708 and 145 535 289, for BF16Cat, Spinn3r and ClueWeb09, respectively. Despite the large number of tags, 22 distinct tags account for 93.79%, 87.96%, and 90.82% of all tag usages in BF16Cat, Spinn3r, and ClueWeb09, respectively. The datasets are profiled according to the frequency of each 22 tags relative to each other, in Figure 1.

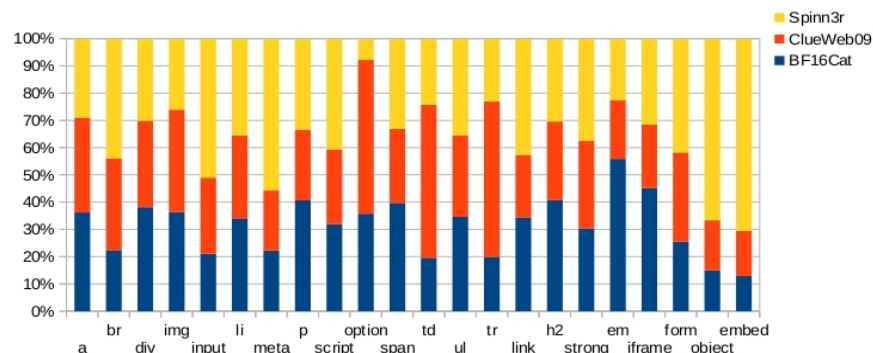


Figure 1. Relative distribution of the 22 most popular HTML tags across datasets.

⁸ National Building Specification website: <http://www.thenbs.com/resources/directory/index.asp>

⁹ Smashing Magazine Showcase of Beautiful Fashion Websites: <http://www.smashingmagazine.com/2009/03/12/showcase-of-beautiful-fashion-websites/>

¹⁰ Politics Resources website: <http://www.politicsresources.net/official.htm>

Platform and File Formats in Use

In this section, we emphasise the differences between web pages across different domains of BF16Cat to draw out the distinctions between different communities. One distinguishing factor comes from the content management platform (e.g. WordPress¹¹ and/or Joomla¹²) the community chooses to adopt, and another from types of embedded content (e.g. file formats).

Communities Profiled by Content Management Systems They Use

The HTML code for the home pages belonging to the categories in BF16Cat have been analysed to produce the profile of each community in terms of the number of times each content management system (CMS) was adopted for the community home page. The results are displayed in Figure 2.

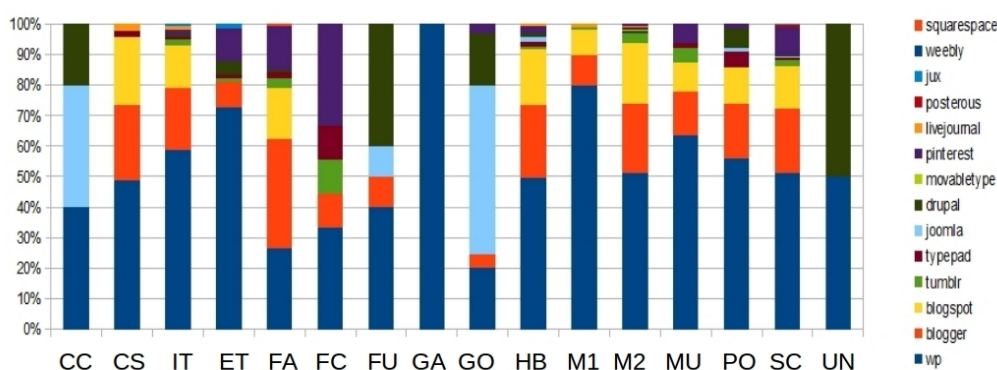


Figure 2. Content management systems (colour coded) preferred by each community (x-axis, labels are from Table 2).

While Figure 2 shows many platforms being used by each community, it also suggests each community has a fairly strong preference for one or two platforms over others. For example both game blogs and mathematics blogs are predominantly deployed on WordPress. The figure illustrates substantial amount of variation across communities (e.g. government websites seem to have a strong preference for Joomla).

Indeed, in Figure 3, a hierarchical clustering of communities is displayed (based on Pearson coefficient (Pearson, 1895) distance measure and Ward’s method (Ward, 1963) for hierarchical clustering) which illustrates a clear division between the blogging websites (bottom sub-tree in pink) and non-blogging websites (top sub-tree in blue). The clusters include predictable pairings, for example, of funding councils and universities, of game blogs and entertainment blogs, of computing blogs and mathematics blogs, and of fashion blog and fashion company websites.

There are some curious connections that also emerged (e.g. mathematics I and music; health blogs and mathematics II) but further investigation is required to make any conclusions. Nevertheless, the clustering results suggest that platform selection might be influenced by the community of the blogger.

¹¹ WordPress: <http://wordpress.com/>

¹² Joomla: <http://www.joomla.org/>

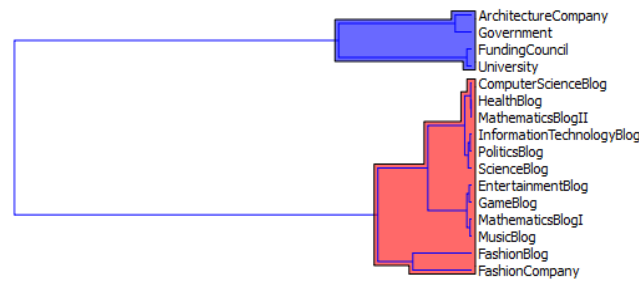


Figure 3. Hierarchical clustering of communities on the basis of preference of content management systems.

Communities Profiled by File Extensions They Use

The HTML can also be examined for “file-extensions”. This information was extracted by extracting patterns consisting of a full stop followed by up to four characters from all the HTML tag attribute values. A hierarchical clustering of domains based on these “extensions” was performed to produce the diagram in Figure 4.

The result here shows that there is a clear division between blogging and non blogging communities. It also shows some obvious connections between computer science and mathematics, and music and entertainment. Some of the earlier clusters that emerged with respect to platform selection re-emerge with respect to “extensions”, which suggests that the clustering might be influenced by “file extensions” associated to content management system attributes rather than content generated by blog/website authors.

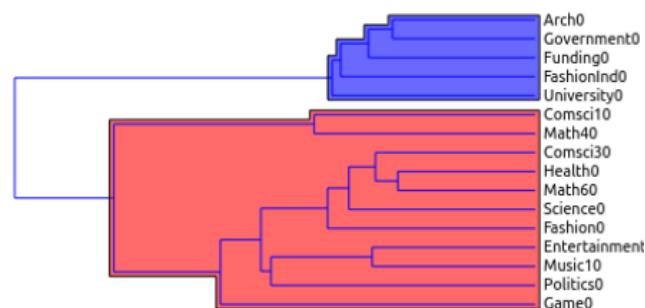


Figure 4. Hierarchical online community clustering with respect to “file extensions” used in the HTML codes of the community.

Web Page Links

As this study was only intended to be a lightweight study, the concepts of betweenness, closeness, cliques, authorities and hubs of social networks have not been attempted. Instead the investigation was limited to a quick comparison of “referencing behaviour” of the communities.

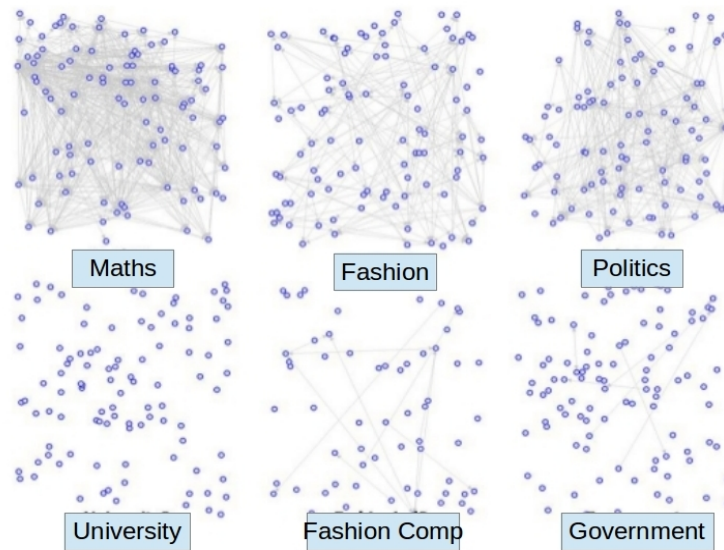


Figure 5. Link structure derived from three blog and three non-blog home pages.

The network traffic can vary considerably across dataset. As an example, in Figure 5, we have included the network traffic of six of the collections in the BF16Cat dataset. Each node in the structure represents a webpage in the collection and the edges represent instances where a page has provided a hyperlink to another blog in the community using the “href” attribute field value for the HTML tag <a>. The representation in the figure is based on 100 random pages from each of the collections. It does not include references to their own websites.

The figure clearly shows the Maths network to be the busiest network among the six. However, more striking is the noticeable difference between blogs and non-blogs with respect to the number of connections made to external sites in the same domain. In contrast to this, the graph in Figure 6 shows that politics blog pages have the most number of links per page, on average.

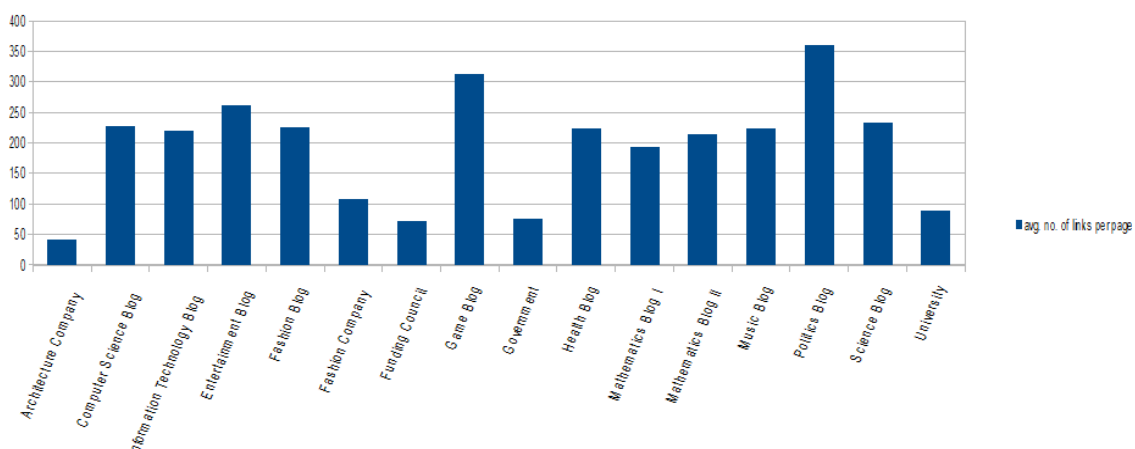


Figure 6. Average number links per URL from each subcategory of BF16Cat.

This suggests that, while mathematics blogs tend to provide references to resources within the community, politics blogs point to resources outside of the politics blogging community. This observation motivates the general examination of average number of links, distinct links, external links and self-referential links. This is presented in Table 3. As might be expected, there seem to be more links on a blog web page on average than on a non-blog web page. However, non-blogs seem to be much more self-referential than blogs, that is they point to resource within their domain much more than to resources elsewhere in the same community (for example, universities tend not to link to other universities).

Table 3. Characteristics of linked references across different communities.

Category	Average number of links per page (total no. links)		Distinct links	No. of non-self referential links	Self-referential links (%)
Construction Company	41.82	(1129)	843	125	0.8892825509
Computer Science	226.76	(9297)	6823	4773	0.4866085834
Information Technology	219.57	(30300)	21493	10313	0.6596369637
Entertainment	261.21	(28733)	19357	9960	0.6533602478
Fashion Blog	225.24	(36940)	28660	23513	0.3634813211
Fashion Company	105.57	(6440)	4926	1154	0.8208074534
Funding Council	71.84	(3664)	2803	503	0.8627183406
Game Blog	312.00	(2184)	1479	714	0.6730769231
Government	75.80	(43356)	31524	8464	0.8047790387
Health Blog	224.42	(29175)	21408	14054	0.5182862039
Mathematics Blog I	195.96	(21360)	15251	8977	0.5797284644
Mathematics Blog II	214.62	(118471)	83283	44349	0.6256552236
Music Blog	223.93	(15675)	11357	8959	0.4284529506
Politics Blog	361.08	(38636)	27709	20163	0.4781292059
Science Blog	233.10	(249652)	155420	129754	0.4802605226
University	89.83	(8983)	7369	2095	0.7667816988

User Generated Category and Topic Tags

User generated tags are largely associated with social networking media technology. The generation of topic tags is expected to be fairly active in the blogging communities but not so much so in non-blogging communities. Indeed there were no user generated categories and/or tags in the non-blog subcategory collections of BF16Cat dataset except for six categories (“tv”, “tourism”, “sports-news”, “somali-politics”, “somali-news”, and “business-news”) that were used in one of the government home pages.

Here we present some observations on how we might combine the analysis of category and topic tags generated by users with graph analysis to produce a

representation of domain knowledge. As a starting point, the harvested blogs were investigated to create term graphs. The steps for creating the graphs are as follows:

1. The home page address is the root node of the graph.
2. Each category and topic tag that appears in the home page is a node of the graph.
3. There is a directed edge from the root node to a category for every category found on the page.
4. There is a directed edge from category to topic tag for each category and topic tag that appear together in a post.

The initial result of taking these steps, a sample graph created from a game blog¹³ and a mathematics blog¹⁴, is presented in Figures 7 and 8.

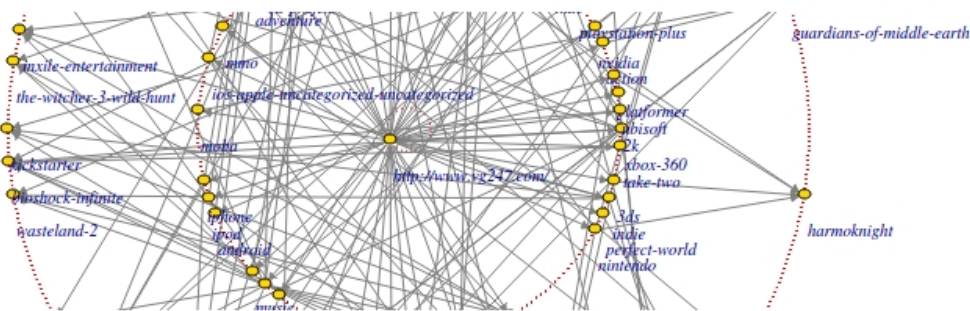


Figure 7. Graph constructed from user generated categories and topic tags for a sample game blog.

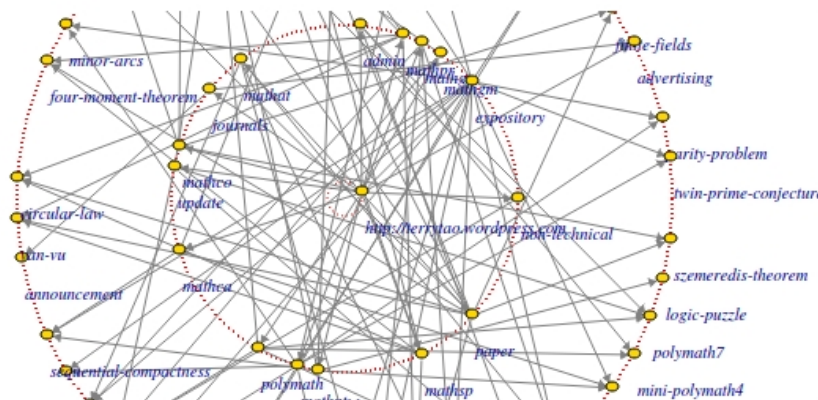


Figure 8. Graph constructed from user generated categories and topic tags for a sample mathematics blog.

The proposal for the next step to refine the graph to reflect community knowledge is to layer the graph according to the URL frequency associated to each category and topical tag terms across the community. For example, the graph for the mathematics blog in Figure 8 shares the category “expository” with at least four other mathematics blogs. The “expository” node leads to 16 other topics, such as “Bolzano-Weierstrass-theorem” and “experimental-sums”. The term “experimental-sums” is also linked from

¹³ VG 24/7: <http://www.vg247.com>

¹⁴ Terry Tao Blog: <http://terrytao.wordpress.com>

the category “polymath” which is not shared as much in the community. This shared topic suggests that there is a relationship between “expository” areas and “polymath”. In fact, Polymath Projects¹⁵ has been established for “massively collaborative online mathematical projects” of an experimental nature.

Benefits to Digital Curation

The topics in this paper highlight the potential of social media infrastructure for sharing information when profiling a target “designated community”. The main objective of the study was to better define “designated communities” by providing a quantifiable profile of communities to complement labour-intensive qualitative processes. The profile could be used to inform repositories right at the beginning of the collection process by 1) adding community context to characterise the target materials to be collected, and 2) providing first steps to assess the risks associated with the variety of adopted technologies and formats, changes in concepts with respect to knowledge organisation, and social impact of information loss given the community knowledge exchange network. Studies of performing retrospective data analytics on collections (e.g. Jackson, 2012; Rimkus et al., 2014) unearth collection standards and trends but do not help to determine whether or not these standards meet the needs of the community and/or expose the technological culture that imposes obstacles to enforcing the use of standards. By studying the community in action as information sharing takes place, this gap can be bridged.

The study presented here was motivated in the development of policies for the preservation of weblogs, and as such the target materials come with web and social media platform infrastructure in place to make community profiling viable. The use of the approach on material from a non-web environment may be, therefore, questioned. However, this can be overcome through two immediate solutions: 1) by embedding the repository within the social media information sharing workflow, and/or 2) by developing a agreed approach for mapping the “designated community” to an online presence (cf. Liu et al., 2012).

Conclusions

In this article we discussed the ways in which communities follow a recognisable pattern with respect to: the content management platform they select to create their websites, the file formats in use within the community, how they reference resources within the community, categories and topics they use, and how they organise these categories and topics. In particular, it has been demonstrated how these practices differ between blogging and non-blogging communities, and between different types of sub-domains associated to these. The article has demonstrated how these practices can be studied using an automated process based on the HTML content of the home pages associated to community website. This could be highly valuable for studying the concept of “designated communities” by supplementing qualitative approaches already used in the curation community. The study highlights the possibility that the maturity of social media and other collaborative technologies may be leveraged to capture designated community practices even when target materials are not created on the web.

¹⁵ Polymath Projects wiki: <http://michaelnielsen.org/polymath1/index.php>

Acknowledgements

The author would like to thank the European Commission for co-funding the BlogForever project (grant no. 269963) and the University of Glasgow for providing the funds and resources for carrying out this research. I would like to thank Seamus Ross for valuable feedback and Matthew Barr for directing me to the relevant game blogs.

References

- Alliance for Permanent Access. (n.d.). *Identifying the designated community*.
<http://www.alliancepermanentaccess.org/index.php/learningunit/identifying-the-designated-community/>
- Coulon, F. (2005). *The use of social network analysis in innovation research: A literature review*. Retrieved from
<http://www.druid.dk/conferences/winter2005/papers/dw2005-305.pdf>
- ISO. (2003). *ISO 14721:2003. Space data and information transfer systems – Open Archival Information System – Reference model*.
- Jackson, A.N. (2012). *Formats over time: Exploring UK web history*. Paper presented at iPres 2012, Toronto, Canada. Retrieved from <http://arxiv.org/abs/1210.1714>
- Lampe, C., Ellison, N.B., & Steinfield, C. (2008). Changes in use and perception of Facebook. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08)* (pp. 721–730). New York, NY: Association for Computing Machinery. doi:10.1145/1460563.1460675
- Liu, X., He, Q., Tian, Y., Lee, W-C., McPherson, J., & Han, J. (2012). Event-based social networks: Linking the online and offline social worlds. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)* (pp. 1032–1040). New York, NY: Association for Computing Machinery. doi:10.1145/2339530.2339693
- Massimi, M., Bender, J.L., Witteman, H.O., & Ahmed, O.H. (2014). Life transitions and online health communities: Reflecting on adoption, use, and disengagement. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '14)* (pp. 1491–1501). New York, NY: Association for Computing Machinery. doi:10.1145/2531602.2531622
- O’Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2013). Uncovering the wider structure of extreme right communities spanning popular online networks. In *Proceedings of the 5th Annual ACM Web Science Conference (WebSci '13)* (pp. 276–285). New York, NY: Association for Computing Machinery. doi:10.1145/2464464.2464495
- Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, 240–242.

- Rimkus, K., Padilla, T., Popp, T., & Martin, G. (2014). Digital preservation file format policies of ARL member libraries: An analysis. *D-Lib Magazine*, 20(3\4).
doi:10.1045/march2014-rimkus
- Rowson, J., Broome, S., & Jones, A. (2010). *Connected communities: How social networks power and sustain the Big Society*. Retrieved from Royal Society for the encouragement of Arts, Manufactures and Commerce website: <https://www.thersa.org/discover/publications-and-articles/reports/connected-communities-how-social-networks-power-and-sustain-the-big-society/>
- Sedghi, A., & Evans, L. (2012, March 15). World's top 100 universities 2013: Their reputations ranked by Times Higher Education [Web log post]. Retrieved from The Guardian, Higher Education Datablog web log: <http://www.guardian.co.uk/news/datablog/2012/mar/15/top-100-universities-times-higher-education>
- Ward, J.H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.
doi:10.1080/01621459.1963.10500845
- York University. (2013). Designated community definition. In *Digital Preservation Policy*. Retrieved from <https://digital.library.yorku.ca/documentation/digital-preservation-designated-community-definition>